Sanity Checks for Saliency Maps

Julius Adebayo *, Justin Gilmer *, Michael Muelly *, Ian Goodfellow *, Moritz Hardt ^ *, Been Kim *

Work was done during the Google AI residency program, *MIT, ^ UC Berkeley, *Google Brain.
Interpretability for Deep CNNs

Input

Inception V3

Explanation

Saliency Map: Gradient

Predictions

Junco Bird

[SVZ'13]
Saliency Maps

Input

Inception V3

Predictions

Junco Bird

Explanation

Gradient SmoothGrad Guided BackProp Guided GradCAM Integrated Gradients Integrated Gradients SmoothGrad Gradient Input

[SVZ’13] [STKVW’17] [SDBR’14] [SCDVPB’17] [STY’17] [STKVW’17] [SGSK’16]

Our Question

Do these methods work?
Sanity Check: **Model Parameter Randomization**

Inception V3 with trained weights

Junco Bird

Guided BackProp Explanation
Sanity Check: **Model Parameter Randomization**

- Inception V3 with trained weights
- Inception V3 with Randomized weights

Guided BackProp Explanation

- Junco Bird
- Wheaten Terrier
Sanity Check: Model Parameter Randomization

Janco Bird

Inception V3 with trained weights

Wheaten Terrier

Inception V3 with Randomized weights

Visually indistinguishable!
Cascading Randomization Inception V3

To assess sensitivity to model parameters, we randomize the parameters of a model and compare explanations derived from a model with random weights to one with trained weights.
To assess the degree to which saliency maps capture the input-label relationship, we compare maps derived from models trained on data with permuted labels to maps derived from models trained with true labels [ZBHRV’17].
Conclusion

1. **Sanity Checks**: model parameter randomization & data randomization tests.

2. **Visual assessment of saliency maps is inadequate.**

3. The inductive bias of a model, e.g. convolution for CNNs, has a strong influence on saliency maps.

4. Nie et. al (ICML 2018), independent of this work, observed similar findings.
Thanks!

Poster #29, Session A: 10:45am - 12:45pm
@Room 210 & 230.